



Multi-agent reinforcement learning for the coordination of residential energy flexibility

Flora Charbonnier
Energy and Power Group, University of Oxford

DTU CEE Summer School 2022 – Advanced Optimization, Learning, and Game-Theoretic Models in Energy Systems

Agenda

- I. Research objectives
- II. Challenges for the coordination of residential energy flexibility
- III. The landscape of distributed energy resources coordination
- IV. Local energy system description
- V. Reinforcement learning methodology
- VI. Results
- VII. Next steps for multi-agent reinforcement learning scalability

Questions

I. Research Objectives

- Increasing need for demand–side flexibility given higher shares of renewable energy
- High potential of residential sector flexibility
 - 55% of electricity, transport and heat energy in the UK
 - 50% share of peak with higher marginal value
 - Increasing resource ownership but so far excluded from demand side response

Can we coordinate residential energy flexibility at scale?

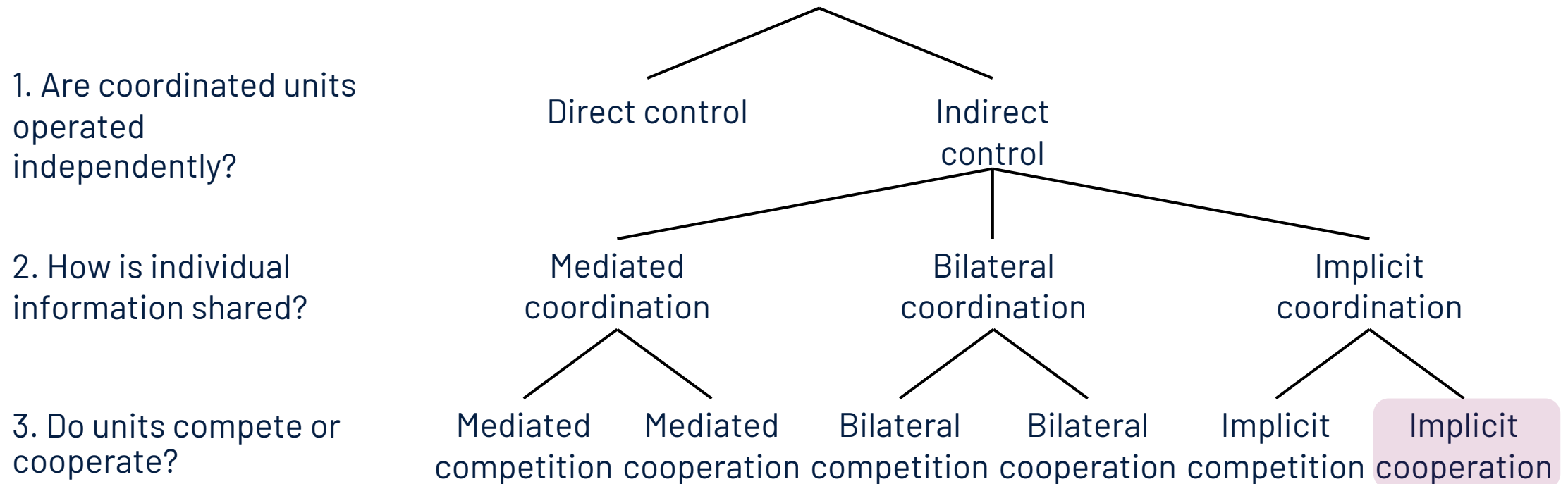
II. Challenges for the coordination of residential energy flexibility

1. Costs
2. Acceptability & privacy
3. Computational feasibility

Can we coordinate residential energy flexibility in a privacy-preserving, cost-effective, scalable way for agents under partial observability in stochastic environments?

III. The landscape of distributed energy resources coordination

Taxonomy of distributed resources coordination



IV. Local energy system description

Maximise $\sum_{\forall t \in \mathcal{T}} F_t = \sum_{\forall t \in \mathcal{T}} -(c_g^t + c_d^t + c_s^t)$

- Grid costs: energy, losses, greenhouse gas emissions

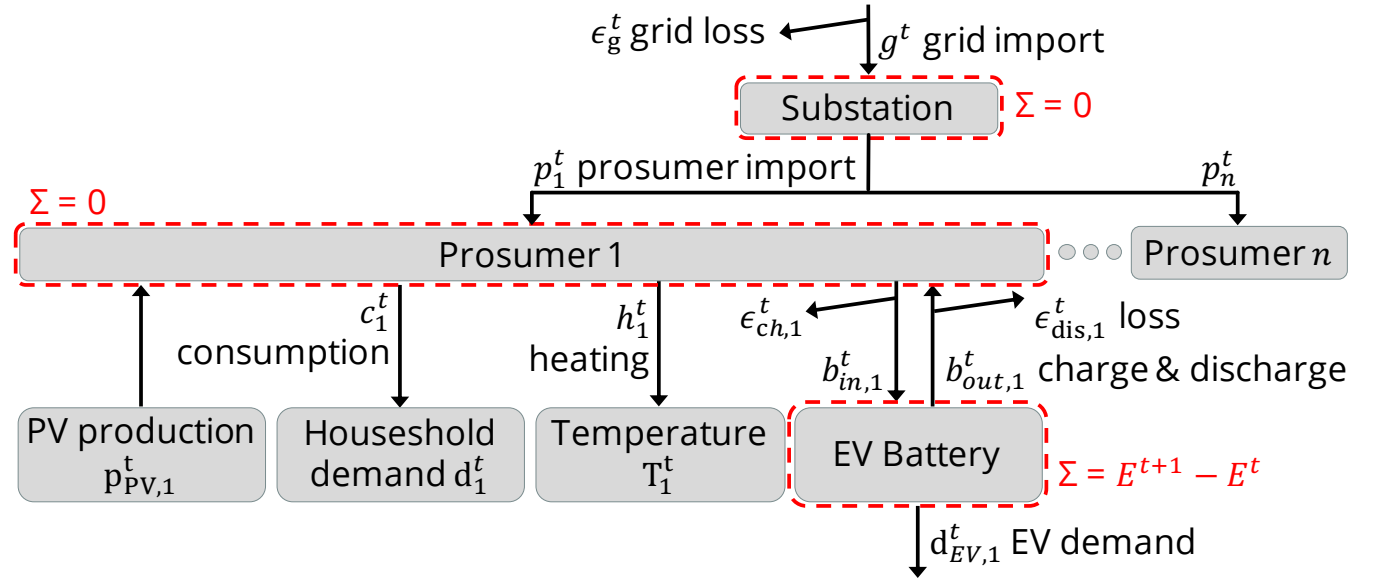
$$c_g^t = C_g^t \left(g^t + \frac{R}{V^2} (g^t)^2 \right)$$

- Distribution costs: exports

$$c_d^t = C_d \sum_{i \in \mathcal{P}} \max(-p_i^t, 0)$$

- Battery costs: depreciation

$$c_s^t = C_s \sum_{i \in \mathcal{P}} (b_{in,i}^t + b_{out,i}^t)$$

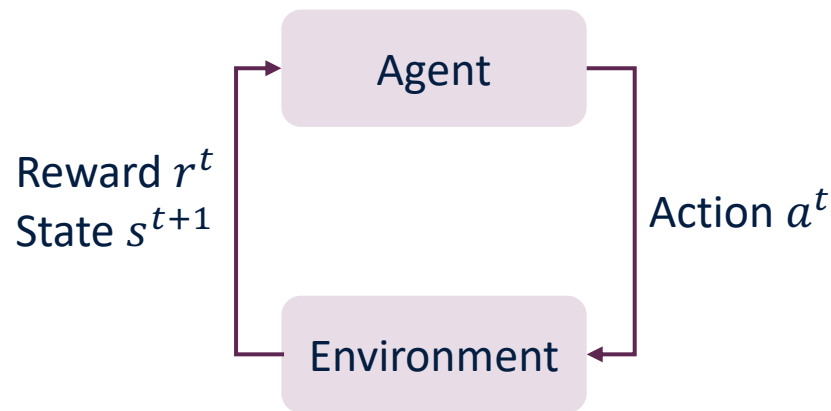


V. Reinforcement learning methodology

Reinforcement learning

- Artificial intelligence framework
- Agent interacts with the environment to learn the set of actions in each state to maximise rewards

Markovian decision process



$$\sum_{\forall t \in \mathcal{N}} r_t = F$$

Q-learning

- Disaggregate global coordination problem into local decisions by each agent at each time step
- A Q-table stores state-actions values representing the expected value of all future rewards

$$Q(s, a) \triangleq E^\pi [r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} \dots | s_t = s, a_t = a]$$

$$\hat{Q}(s, a) \leftarrow \hat{Q}(s, a) + \alpha (r_t + \gamma \hat{V}(s^{\text{next}}) - \hat{Q}(s, a))$$

$$\hat{V}(s) = \max_{a^* \in \mathcal{A}(s)} \hat{Q}(s, a^*)$$

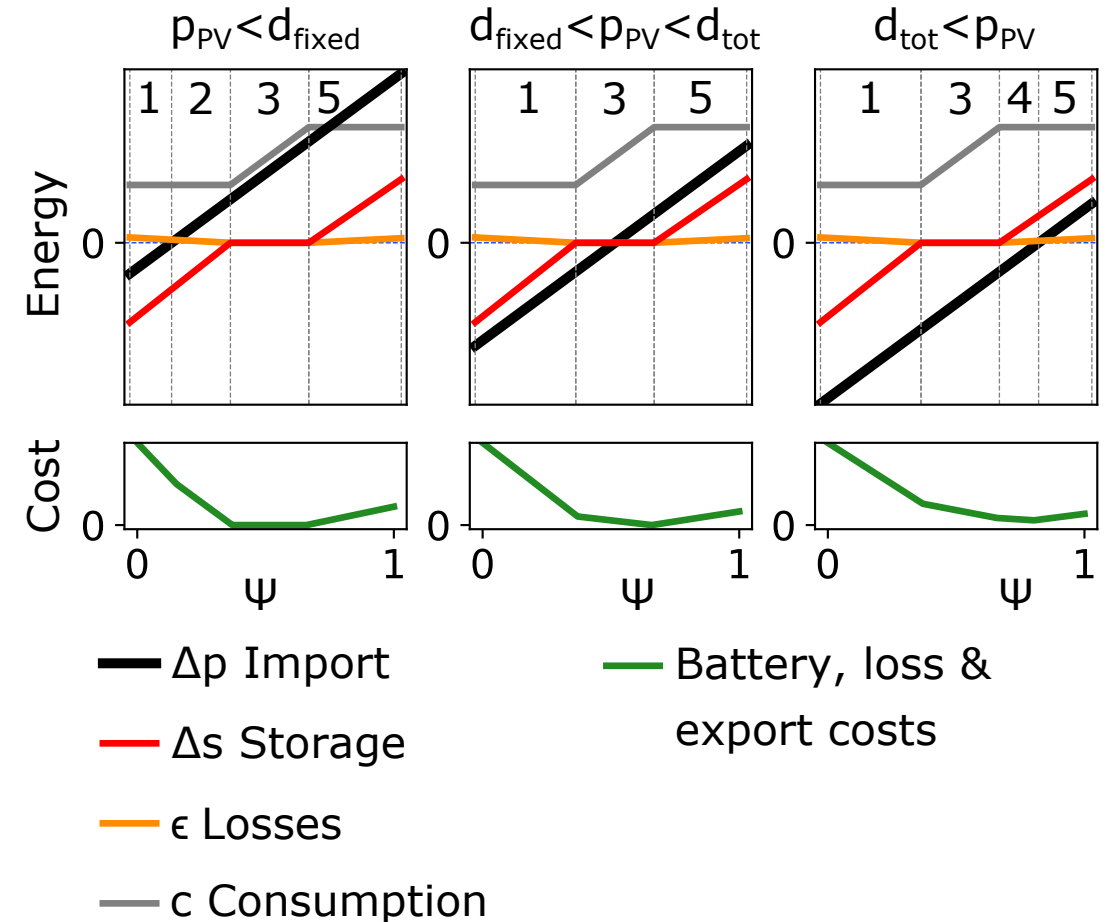
where γ and α are the discount and learning factors

V. Reinforcement learning methodology

Unify decision variables in a unique action dimension Ψ to mitigate curse of dimensionality

Between max. exports $\Psi = 0$ and imports $\Psi = 1$

1. Exporting all to none of the initial energy stored
2. From meeting fixed loads with initial energy stored to importing the required amount
3. Consuming none to all flexible loads
4. Exporting all to storing all additional PV energy
5. Importing no additional energy to battery capacity



V. Reinforcement learning methodology

Variations of the learning method

Update rule: $Q(s_i^t, a_i^t) \leftarrow Q(s_i^t, a_i^t) + \alpha\delta$

Multi-agent learning structures

1. Distributed learning
2. Centralised learning

Experience sources

1. Environment explorations
2. Omniscient convex optimisations

Reward definitions

1. Total reward

$$\delta = r_0^t + \gamma V(s_i^{t+1}) - Q^0(s_i^t, a_i^t)$$

2. Marginal reward

$$d = r_0^t - r_{0, a_i=\text{default}}^t$$

$$\delta = d + \gamma V^{\text{diff}}(s_i^{t+1}) - Q^{\text{diff}}(s_i^t, a_i^t)$$

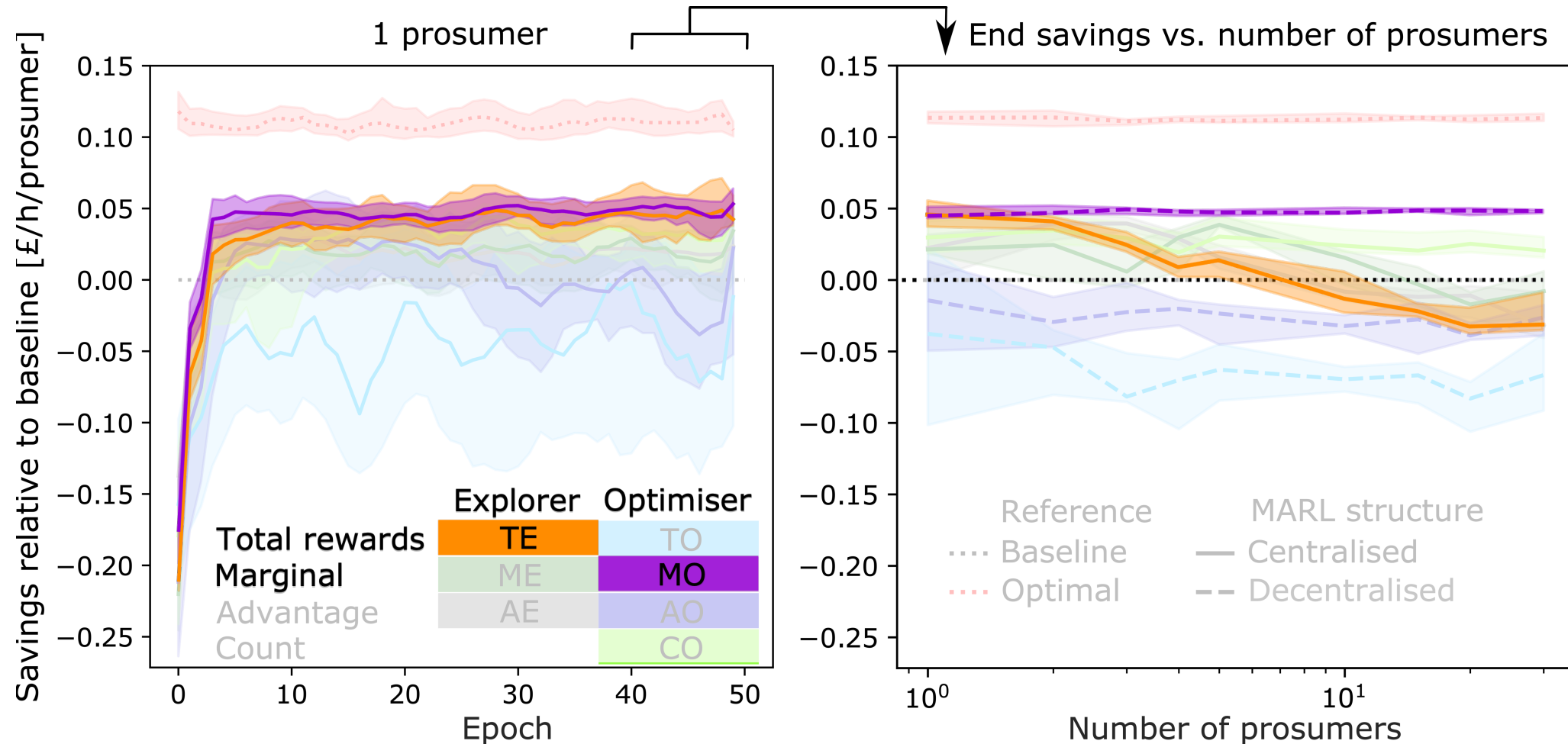
3. Advantage reward

$$\delta = (Q^0(s_i^t, a_i^t) - Q^0(s_i^t, a_{i=\text{default}}^t)) - Q^A(s_i^t, a_i^t)$$

4. Count

$$\alpha\delta = 1$$

VI. Results

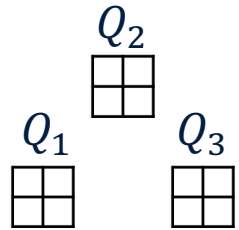


VII. Next step: scalable MARL

Remaining limitations to scalability due to optimisation and baselining during training

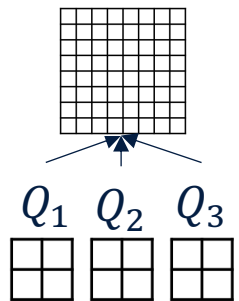
- Pre-learning is not time-critical and this is not a problem during implementation
- But can we find less computationally intensive equilibrium mechanisms during training?

VII. Next step: scalable MARL



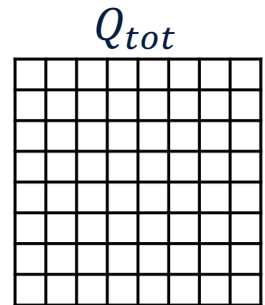
Independent Q-learning

The simplest option: No centralised action-value function, learn individual action-value function Q_a individually



Factored value estimator

Value-based methods in between the two extremes, learn a centralised but factored Q_{tot}



Centralised value estimator

Actor-critic methods with a fully centralised state-action value function Q_{tot} used to guide the optimisation of decentralised policies

Conclusion

Developed a cost-effective, scalable methodology suited to residential energy flexibility coordination, overcoming three main hurdles for their integration:

1. Costs → minimal communication infrastructure
2. Acceptability → no trade-offs in prosumers comfort, no sharing of personal data
3. Computations → statistical approach, scalable local decision-making

Questions

Flora Charbonnier

Energy and Power Group

Department of Engineering Science

flora.charbonnier@pmb.ox.ac.uk

